Looking Through the Glass: Neural Surface Reconstruction Against High Specular Reflections

Jiaxiong Qiu[1] Peng-Tao Jiang[2] Yifan Zhu[1] Ze-Xin Yin[1] Ming-Ming Cheng[1] Bo Ren[1*]
[1]VCIP, CS, Nankai University [2]Zhejiang University

Abstract

Neural implicit methods have achieved high-quality 3D object surfaces under slight specular highlights. However, high specular reflections (HSR) often appear in front of target objects when we capture them through glasses. The complex ambiguity in these scenes violates the multi-view consistency, then makes it challenging for recent methods to reconstruct target objects correctly. To remedy this issue, we present a novel surface reconstruction framework, NeuS-HSR, based on implicit neural rendering. In NeuS-HSR, the object surface is parameterized as an implicit signed distance function (SDF). To reduce the interference of HSR, we propose decomposing the rendered image into two appearances: the target object and the auxiliary plane. We design a novel auxiliary plane module by combining physical assumptions and neural networks to generate the auxiliary plane appearance. Extensive experiments on synthetic and real-world datasets demonstrate that NeuS-HSR outperforms state-of-the-art approaches for accurate and robust target surface reconstruction against HSR. Code is available at https://github.com/JiaxiongQ/NeuS-HSR.

1. Introduction

Reconstructing 3D object surfaces from multi-view images is a challenging task in computer vision and graphics. Recently, NeuS [45] combines the surface rendering [3, 12, 35, 52] and volume rendering [8, 29], for reconstructing objects with thin structures and achieves good performance on the input with slight specular reflections. However, when processing the scenes under high specular reflections (HSR), NeuS fails to recover the target object surfaces, as shown in the second row of Fig. 1. High specular reflections are ubiquitous when we use a camera to capture the target object through glasses. As shown in the first row of Fig. 1, in the captured views with HSR, we can recognize the virtual image in front of the target object. The virtual image introduces the photometric variation on the object surface visually, which degrades the multi-view consistency and encodes extreme ambiguities for rendering, then confuses NeuS to reconstruct the reflected objects instead of the target object.

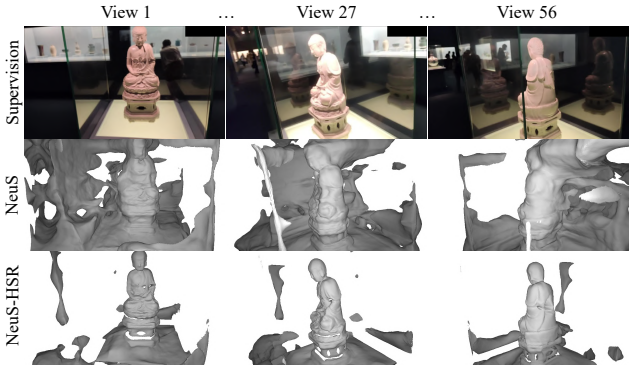


Figure 1. 3D object surface reconstruction under high specular reflections (HSR). Top: A real-world scene captured by a mobile phone. Middle: The state-of-the-art method NeuS [45] fails to reconstruct the target object (*i.e.*, the Buddha). Bottom: We propose NeuS-HSR, which recovers a more accurate target object surface than NeuS.

To adapt to the HSR scenes, one intuitive solution is firstly applying reflection removal methods to reduce HSR, then reconstructing the target object with the enhanced target object appearance as the supervision. However, most recent single-image reflection removal works [4, 9, 23, 24, 26, 40] need the ground-truth background or reflection as supervision, which is hard to be acquired. Furthermore, for these reflection removal methods, testing scenes should be present in the training sets, which limits their generalization. These facts demonstrate that explicitly using the reflection removal methods to enhance the target object appearance is impractical. A recent unsupervised reflection removal approach, NeRFReN [18] decomposes the rendered image into reflected and transmitted parts by implicit representations. However, it is limited by constrained view directions and simple planar reflectors. When we apply it to scenes for multi-view reconstruction, as Fig. 3 presents, it takes the target object as the content in the reflected image

*Bo Ren is the corresponding author.

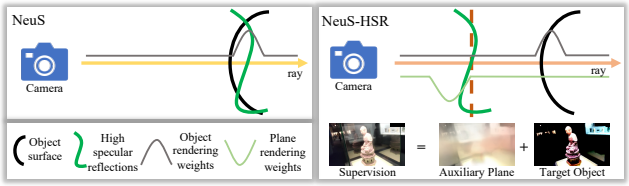


Figure 2. NeuS-HSR. High specular reflections (HSR) make NeuS tend to reconstruct the reflected object in HSR. NeuS-HSR physically decomposes the rendered image into the target object and auxiliary plane parts, which encourages NeuS to focus on the target object.

and fails to generate the correct transmitted image for target object recovery.

The two-stage intuitive solution struggles in our task as discussed above. To tackle this issue, we consider a more effective decomposition strategy than NeRFReN, to enhance the target object appearance for accurate surface reconstruction in one stage. To achieve our goal, we construct the following assumptions:

Assumption 1 *A scene that suffers from HSR can be decomposed into the target object and planar reflector components. Except for the target object, HSR and most other contents in a view are reflected and transmitted through the planar reflectors (*i.e.*, glasses).*

Assumption 2 *Planar reflectors intersect with the camera view direction since all view direction vectors generally point to the target object and pass through planar reflectors.*

Based on the above physical assumptions, we propose NeuS-HSR, a novel object reconstruction framework to recover the target object surface against HSR from a set of RGB images. For **Assumption** 1, as Fig. 2 shows, we design an auxiliary plane to represent the planar reflector since we aim to enhance the target object appearance through it. With the aid of the auxiliary plane, we faithfully separate the target object and auxiliary plane parts from the supervision. For the target object part, we follow NeuS [45] to generate the target object appearance. For the auxiliary plane part, we design an auxiliary plane module with the view direction as the input for **Assumption** 2, by utilizing neural networks to generate attributes (including the normal and position) of the view-dependent auxiliary plane. When the auxiliary plane is determined, we acquire the auxiliary plane appearance based on the reflection transformation [16] and neural networks. Finally, we add two appearances and then obtain the rendered image, which is supervised by the captured image for one-stage training.

We conduct a series of experiments to evaluate NeuS-HSR. The experiments demonstrate that NeuS-HSR is superior to other state-of-the-art methods on the synthetic dataset and recovers high-quality target objects from HSR-effect images in real-world scenes.

Figure 3. Decomposition of NeRFReN [18]. NeRFReN fails to separate specular reflections and the target object appearance in this view, then makes NeuS fail to recover the target object surface.

To summarize, our main contributions are as follows:

- We propose to recover the target object surface, which suffers from HSR, by separating the target object and auxiliary plane parts of the scene.
- We design an auxiliary plane module to generate the appearance of the auxiliary plane part physically to enhance the appearance of the target object part.
- Extensive experiments on synthetic and real-world scenes demonstrate that our method reconstructs more accurate target objects than other state-of-the-art methods quantitatively and qualitatively.

2. Related Works

2.1. Traditional Surface Reconstruction

The classical multi-view surface reconstruction methods mainly consist of two categories: photometric stereo [5, 6, 19, 50] and multi-view stereo [11, 13–15, 36–38] reconstruction. The photometric stereo reconstruction is limited by the strict experimental environment. For the multi-view stereo reconstruction, the input images are captured around the target object in common scenes. The early multi-view stereo methods [11, 15, 36, 37] focus on the object surfaces with diffuse materials. They all obey the Lambertian assumption that the same detected region of the object surfaces has little change in all views. However, obvious specular reflections often occur on objects in real-world scenes, *e.g.*, the highlight. The Lambertian assumption no longer holds in real-world scenes with obvious specular reflections. The widely-used Structure From Motion (SFM) methods [34, 42, 49] is designed to calibrate the camera and produce a sparse depth map at each viewpoint firstly. Then the object surface can be acquired by Poisson Surface Reconstruction [22] with depth fusion. However, the quality of the output surface is easily affected by the feature point detection, and surface areas without rich textures on the target object always have artifacts or empty holes. In this work, we focus on the neural implicit method to achieve accurate 3D object surfaces in more realistic scenarios (*i.e.*, non-Lambertian surfaces).

2.2. Neural Implicit Surface Rendering

Implicit representations based on neural networks have achieved promising results on novel view synthesis [25, 27,

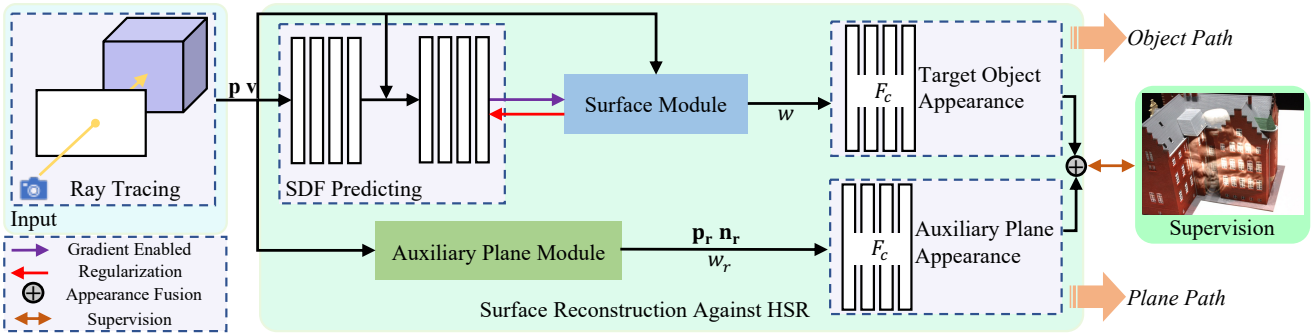


Figure 4. Pipeline of NeuS-HSR. The sampled points $\mathbf{p}$ and the view direction $\mathbf{v}$ are fed into the target object path and the plane path respectively. In the object path, the implicit SDF f is generated by the head neural networks. The surface module takes f, $\mathbf{p}$, and $\mathbf{v}$ as the input, producing the rendering weights w. In the plane path, the reflection module generates the plane normal $\mathbf{n}_r$, 3D locations $\mathbf{p}_r$, and the rendering weights w_r of the auxiliary plane from $\mathbf{p}$ and $\mathbf{v}$. Finally, we acquire two appearances by the appearance function F_c and volume rendering.

29, 39, 47] and 3D reconstruction [7, 10, 30, 31, 33, 41, 44–46, 48, 51, 52]. They have characteristics that classical methods do not have, including flexible resolution and natural global consistency. Surface rendering based on the differentiable ray casting is applied for surface reconstruction with different forms of implicit shape representations, such as the occupancy function [32] and signed distance function (SDF) [52]. IDR [52] extracts surface points on the object surface with the zero-level set of SDF representations, and utilizes neural network gradients to solve a differentiable rendering formulation. UNISURF [31], VolSDF [51] and NeuS [45] learn the implicit surface function by introducing the volume rendering scheme [29], to improve the surface reconstruction quality from captured images. NeuralWarp [7] is a two-stage method for refining the basic model (*e.g.*, VolSDF). NeRS [53] focuses on learning the appearance of object surfaces by introducing the Phong model [20, 21, 43]. It uses a canonical sphere to represent the object surface and learns the object texture with prerequisite masks from a sparse set of images, but it mainly deals with objects with reflective surfaces and produces the object surface without fine details. In contrast with these works, we propose to extend the object surface reconstruction to more challenging HSR scenes in one stage. We aim to correctly recover the object surface through glasses instead of the reflective surface. Our method achieves much better reconstruction accuracy and robustness than previous works in HSR scenes.

3. Method

In this work, we focus on reconstructing the object surfaces in high specular reflection (HSR) scenes. As mentioned in the introduction section, HSR encodes non-target object information, resulting in a low-quality target object surface. To tackle HSR scenes, we introduce a novel object surface reconstruction method, NeuS-HSR, which is based on the implicit neural rendering. The pipeline of NeuS-HSR is shown in Fig. 4.

Specifically, we decompose an HSR scene into two components: the target object and the auxiliary plane. To render the target object appearance, we adopt the scheme of NeuS and pack it as a surface module. To render the auxiliary plane appearance, we design an auxiliary plane module based on the reflection transformation [16] and multi-layer perceptrons (MLPs). Finally, we apply a linear summation to fuse two appearances to obtain the rendered image, which receives supervision from the captured image in a view. In the following, we introduce NeuS-HSR in three parts, including the surface module (Sec. 3.1), the auxiliary plane module (Sec. 3.2), and the rendering process (Sec. 3.3).

3.1. Surface Module

We apply NeuS [45] to render the target object appearance. Specifically, NeuS builds an unbiased and occlusion-aware weight function w based on the implicit SDF $f : \mathbb{R}^3 \to \mathbb{R}$ on each camera ray $\mathbf{h}_s$. Firstly, w is defined as:

$$w(t) = T(t)\rho(t), T(t) = \exp\left(-\int_0^t \rho(u)du\right). \quad (1)$$

where $t \in \mathbb{R}$ is the depth value along $\mathbf{h}_s$, then $\rho(t)$ is constructed by:

$$\rho(t) = \max\left(\frac{-\frac{d\Theta_s}{dt}(f(\mathbf{p}(t)))}{\Theta_s(f(\mathbf{p}(t)))}, 0\right). \quad (2)$$

where the object surface S can be modeled by a zero-level set of the signed distance at the point $\mathbf{p}$: $S = \{\mathbf{p} \in \mathbb{R}^3 | f(\mathbf{p}) = 0\}$. The logistic density distribution $\theta_S(\mathbf{p}) = se^{-s\mathbf{p}}/(1+e^{-s\mathbf{p}})^2$, which is the derivative of the Sigmoid function $\Theta_s(\mathbf{p}) = (1+e^{-s\mathbf{p}})^{-1}$. $1/s$ is the standard deviation of $\theta_S(\mathbf{p})$.

The construction of w is the key contribution of NeuS. It connects the implicit SDF and the volume rendering properly to handle complex object structures. The camera ray $\mathbf{h}_s$ at point $\mathbf{p}$ can be denoted as: $\mathbf{h}_s(t) = \mathbf{o} + t\mathbf{v}$, where $\mathbf{o}$ and $\mathbf{v}$ represent the camera center and view direction sepa-

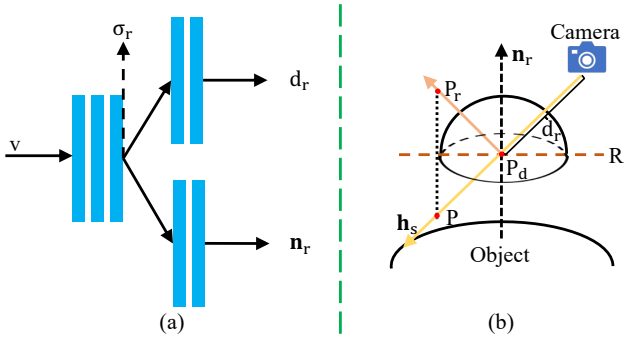


Figure 5. Auxiliary plane module. (a): We design a novel neural network F_r to predict the volume density σ_r, position $\mathrm{d_r}$ and surface normal $\mathbf{n}_r$ of the auxiliary plane R on the camera ray $\mathbf{h}_s$ from the input view direction $\mathbf{v}$. (b): When the auxiliary plane is determined, we project the sampled point P behind R to its reflected point $\mathrm{P_r}$ by the reflection transformation [16].

rately. We sample m points along $\mathbf{h}_s$, then the pixel color value C is acquired by follows:

$$C = \sum_{i=1}^{m} w_i c_i. \tag{3}$$

where $w_i = \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j)$, c_i is the learned color from MLPs, and α_i is the discretization of Eqn. (2).

In our task, C encodes the content of HSR, which causes the ambiguity of w. This ambiguity makes MLPs of SDF predicting tend to model the content of HSR, producing excessive noise around the target object surface. Thus, we propose an auxiliary plane module to divert the attention of MLPs to the target object to handle the interference of HSR.

3.2. Auxiliary Plane Module

In HSR scenes, the virtual images which appear on the planar reflectors and in front of the target object encode extremely ambiguous information for the target object reconstruction. Decomposing the rendered image for enhancing the target object appearance without prior information is an ill-posed problem. NeRFReN [18] applies a depth smoothness prior and a bidirectional depth consistency constraint, to split the rendered image into two components: the transmitted image and the reflected image by implicit representations. This scheme works well in scenes with limited view directions and simple planar reflectors. However, it fails to preserve the target object in transmitted images in HSR scenes. Motivated by NeRFReN, we propose an auxiliary plane module to enhance the target object appearance in the rendered image.

Formally, we use an auxiliary plane R to represent the actual planar reflector for each camera ray. To determine R physically, we design a novel neural network $F_r : \mathbb{S}^2 \rightarrow \mathbb{R}\times\mathbb{R}\times\mathbb{R}^3$ as shown in Fig. 5 (a). F_r maps the view direction $\mathbf{v}$ to the volume density σ_r for generating the rendering weights, and attributes of R (including the position $\mathrm{d_r}$ and the plane normal $\mathbf{n}_r$), that is:

$$\{\sigma_\mathrm{r}, \mathrm{d_r}, \mathbf{n}_r\} = F_r(\mathbf{v}). \tag{4}$$

We assume R is built in the camera coordinate system, the 3D point $\mathrm{p_d} = \mathrm{d_r}\mathbf{v}$ is the interaction of R and $\mathbf{h}_s$. Then $\mathrm{p_d}$ is on R obviously. Given $\mathbf{n}_r = [A, B, C]$, R can be defined as:

$$\mathrm{Ax} + \mathrm{By} + \mathrm{Cz} + \mathrm{D} = 0. \tag{5}$$

where $\mathrm{A}^2 + \mathrm{B}^2 + \mathrm{C}^2 = 1$. We substitute $\mathrm{P_d}$ into Eqn. (5) then have $\mathrm{D} = -\mathrm{d_r}\mathbf{n}_r \cdot \mathbf{v}$.

Moreover, the sampled points along camera rays are part of the inputs for acquiring color values by MLPs. To further model HSR physically, as shown in Fig. 5 (b), for a point p sampled along $\mathbf{h}_s$ and behind R, we project it to the reflected point $\mathrm{p_r}$ along the incident light path based on the reflection transformation [16]. Then MLPs can implicitly trace the incident light to render HSR. Fig. 6 demonstrates the effectiveness of this operation in reducing the interference of HSR. Reflected points help MLPs physically model HSR, then reduce the ambiguity of the scene and recover more accurate target object surfaces. The details of our projection algorithm are explained in supplementary materials.

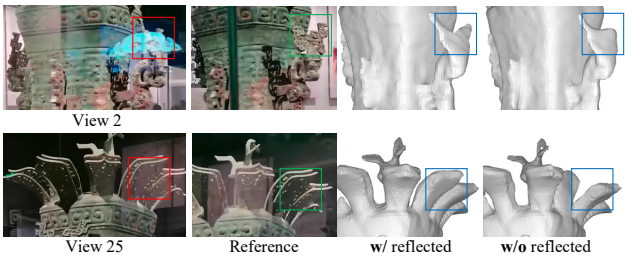


Figure 6. Effect of projecting sampled points to reflected points. Red boxes show the regions of HSR. Green boxes show the reference regions related to red boxes. Blue boxes show effects of HSR on object surface reconstruction.

3.3. Rendering

We adopt the neural network F_c for predicting color values of the object path c_t and the plane path c_a separately. The input of each path is different. For the object path, we follow NeuS and utilize the sampled points $\mathbf{p}$ along the camera ray, surface normal $\mathbf{n}$ of the target object, the view direction $\mathbf{v}$ and features $\mathrm{f_p}$ of the implicit SDF as input. Then we have $c_t = F_c(\mathbf{p}, \mathbf{n}, \mathbf{v}, \mathrm{f_p})$. For the plane path, the sampled points in the camera coordinate system are formed as: $\mathbf{p}' = \mathbf{p} - \mathbf{o}$. As Fig. 7 shows, we utilize both the part points $\mathbf{p}_t$ of $\mathbf{p}'$ in front of R and the reflected points $\mathbf{p}_a$ as the input points $\mathbf{p}_r = \mathbf{p}_t \cup \mathbf{p}_a$. We utilize the plane normal $\mathbf{n}_r$ as input normal. Then for the plane path, we have $c_r = F_c(\mathbf{p}_r, \mathbf{n}_r, \mathbf{v}, \mathrm{f_p})$.

To generate the rendered appearance of each path, we also need to construct two weights of rendering. For the object path, We follow the scheme of NeuS to produce weights w as defined in 3.1. For the auxiliary plane path, given the

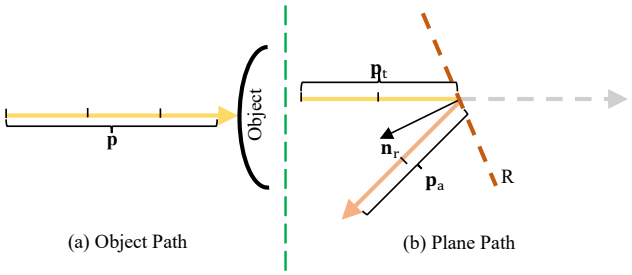


Figure 7. Spatial sampled points along camera rays are used for acquiring color values of two paths. For the object path, we use the original sampled points $\mathbf{p}$. $\mathbf{p}'$ are $\mathbf{p}$ in the camera coordinate system. For the plane path, we use the part points $\mathbf{p}_t$ of $\mathbf{p}'$ which are in front of R along the camera ray and reflected points $\mathbf{p}_a$.

volume density σ_{r} learned from the plane network F_r, we adopt the scheme of NeRFReN to generate weights w_r by:

$$w_r^i = \exp(-\sum_{j=1}^{i-1} \sigma_r^j \delta_j)(1 - \exp(-\sigma_r^i \delta_i)). \tag{6}$$

where $\delta_i = t_{i+1} - t_i$. Finally, the target object appearance $C_t(w, c_t)$ and auxiliary plane appearance $C_r(w_r, c_r)$ can be generated by Eqn. (3). The final rendered image C is obtained by a linear combination of C_t and C_r, which is formulated by:

$$C = \varphi_1 C_t + \varphi_2 C_r. \tag{7}$$

where $\varphi_1 + \varphi_2 = 1$. In practice, we set $\varphi_1 = 0.3$ and $\varphi_2 = 0.7$ by default. The details of this setting are illustrated in the supplementary material.

3.4. Loss Function

During the training procedure of NeuS-HSR, we optimize the difference between the rendered image C and the captured image $\tilde{C}$. We follow the loss function defined in NeuS, which consists of three terms: the color loss $\mathcal{L}_c$ [45,52], the regularized loss $\mathcal{L}_r$ [17] of the implicit SDF and $\mathcal{L}_n$ of the plane normal. The loss functions are formulated as follows:

$$\begin{cases} \mathcal{L}_c = \frac{1}{b} \sum_i \mathcal{L}_1(C_i, \tilde{C}_i), \\ \mathcal{L}_r = \frac{1}{bm} \sum_{k,i} (|\nabla f(\mathbf{p}_k^i)| - 1)^2, \\ \mathcal{L}_n = \frac{1}{b} \sum_i (|\mathbf{n}_r^i| - 1)^2, \end{cases} \tag{8}$$

where b denotes the batch size and m denotes the number of sampled points along a camera ray. Then the final loss function can be defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{c}} + \lambda_1 (\mathcal{L}_{\mathrm{r}} + \mathcal{L}_{\mathrm{n}}). \tag{9}$$

where λ_1 is a constant. Practically, we set $\lambda_1 = 0.1$ by default.

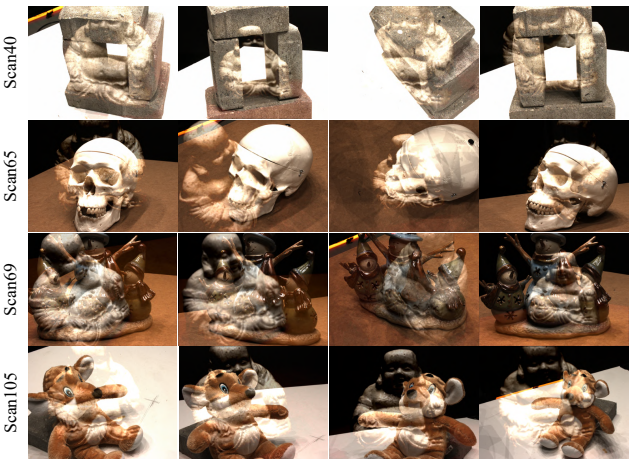


Figure 8. Examples of the synthetic dataset. We apply the widely-used method [55] to synthesize HSR scenes on the DTU dataset [1].

4. Experiments

We conduct extensive experiments which show our method outperforms other approaches quantitatively (Tab. 1) and qualitatively (Fig. 9, Fig. 10). We also provide several ablation experiments to reveal the necessity of our design choices (Fig. 11).

4.1. Datasets

Synthetic dataset. To evaluate the performance of NeuS-HSR and other methods quantitatively, we synthesize 10 scenes from the DTU dataset [1]. We follow the common single-image reflection synthesis method [55] to generate the synthetic dataset. Given the transmission image T (*i.e.*, the image which contains the target object) and the reflection image R′, the image I with reflections can be defined as:

$$\mathrm{I} = \mathrm{T} + \mathcal{K} \otimes \mathrm{R}'. \tag{10}$$

$\mathcal{K}$ is a Gaussian kernel, and $\otimes$ means the convolutional operation. We randomly select a scene as the reflection part, and other scenes are set as the transmission part. Then we adopt Eqn. (10) to acquire HSR scenes. Examples of the synthetic dataset are shown in Fig. 8.

Real-world dataset. To validate the effectiveness of our method in real-world scenes, we collect 6 HSR scenes from the Internet. We utilize the widely-used tool, COLMAP [36], to estimate camera parameters.

4.2. Settings

Implementation details. The signed distance function (SDF) f is parameterized by MLPs, which consists of 8 linear layers. Then the target object surface is produced from the implicit SDF by marching cubes [28]. The auxiliary plane function F_r consists of 3-layer MLPs for pre-

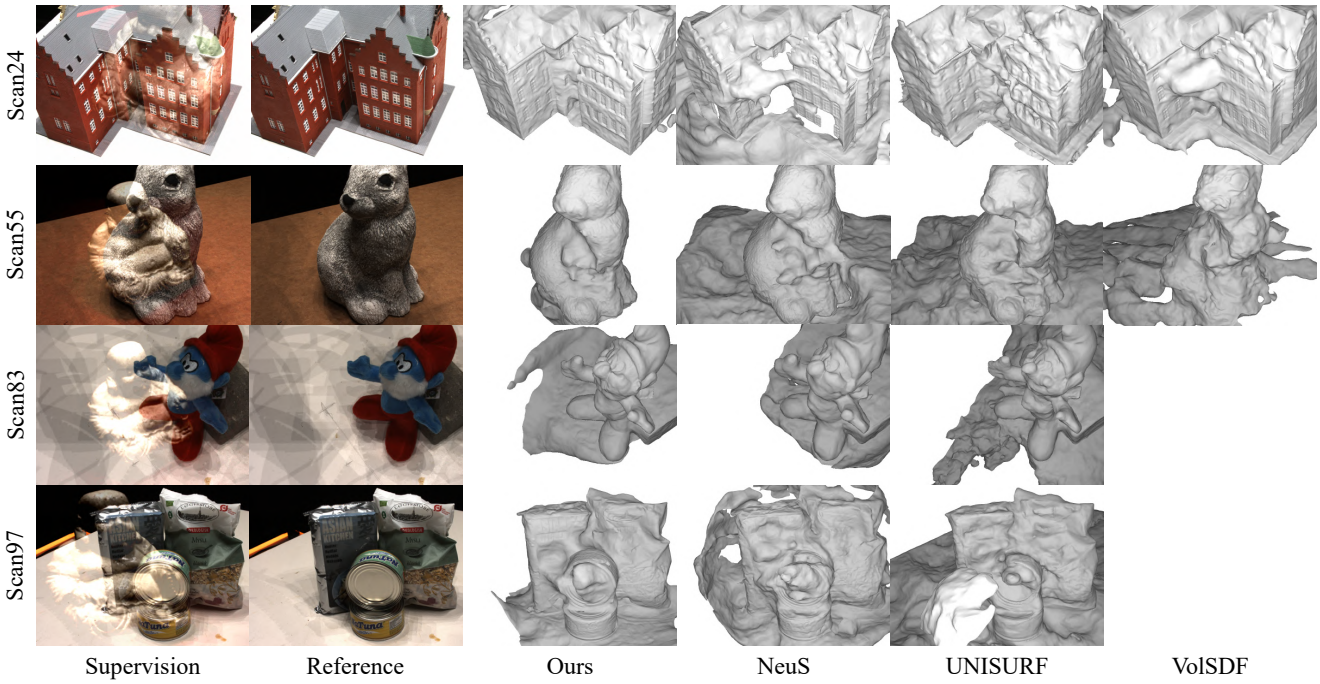


Figure 9. Qualitative comparisons on the synthetic dataset. 'Reference' contains the original appearance of the target object. VolSDF [51] fails to recover the meaningful object surfaces on 'Scan83' and 'Scan97'. Our method achieves better object surface quality when compared to other methods.

Table 1. Quantitative results on the synthetic dataset by measuring the Chamfer distance. VolSDF [51] fails to recover the meaningful object surfaces in the last 7 scenes. Our method outperforms other methods on average. The best metrics are **highlighted**.

ScanID	UNISURF [31]	VolSDF [51]	NeuS [45]	Ours
scan24	2.92	3.89	5.30	**2.07**
scan37	4.26	2.91	2.29	**1.89**
scan40	3.36	2.44	**2.02**	2.17
scan55	2.11	3.95	1.73	**1.25**
scan63	2.73	-	2.75	**1.94**
scan65	1.57	-	**0.93**	1.15
scan69	5.00	-	4.15	**3.54**
scan83	1.81	-	2.55	**1.42**
scan97	3.85	-	4.62	**2.82**
scan105	2.01	-	1.53	**1.31**
mean	2.96	/	2.79	**1.96**

dicting the volume density and 2-layer MLPs for predicting the plane attributes. The rendering appearance function F_c is modeled by 4-layer MLPs. All spatial points are sampled inside a unit sphere, where the scene outside is produced by NeRF++ [54]. Positional encoding [29] is adopted to sampled points $\mathbf{p}$ along camera rays and view directions $\mathbf{v}$. The approximate SDF is pre-processed by the geometric initialization [2]. The batch size of rays is set to 512. We train NeuS-HSR for 200k iterations, consuming about 12 hours on a single NVIDIA Tesla V100 GPU.

Compared Methods. We compare our method against other related approaches with fair settings. The related approaches includes (i) state-of-the-art neural implicit surface reconstruction approaches: NeuS [45], VolSDF [51] and UNISURF [31], (ii) the classical multi-view stereo method: COLMAP [36]. For COLMAP, we apply Screened Poisson [22] to reconstruct its dense mesh from the estimated point cloud. All learning-based models in this paper are trained without ground-truth masks.

4.3. Quantitative Comparisons

For the quantitative evaluation, we conduct comparisons on the synthetic dataset. Following [31, 45, 51], we utilize the Chamfer distance as the evaluation metric, which represents the reconstruction quality of the target object. We report the metric scores in Tab. 1.

4.4. Qualitative Comparisons

As shown in Fig. 9, we present some reconstruction results generated from different methods. It can be observed that other neural implicit methods generate incomplete object surfaces with noise and tends to model the fake specular reflection attached to the target object. The HSR is harmful to these methods to recover the target geometry. On the contrary, our approach achieves clearer results and reconstructs object surfaces with correct geometric details. This fact demonstrates the proposed auxiliary plane module can reduce the interface of the HSR and reconstruct the correct target object surface. Besides, we further evaluate the robustness of each method in more challenging real-world

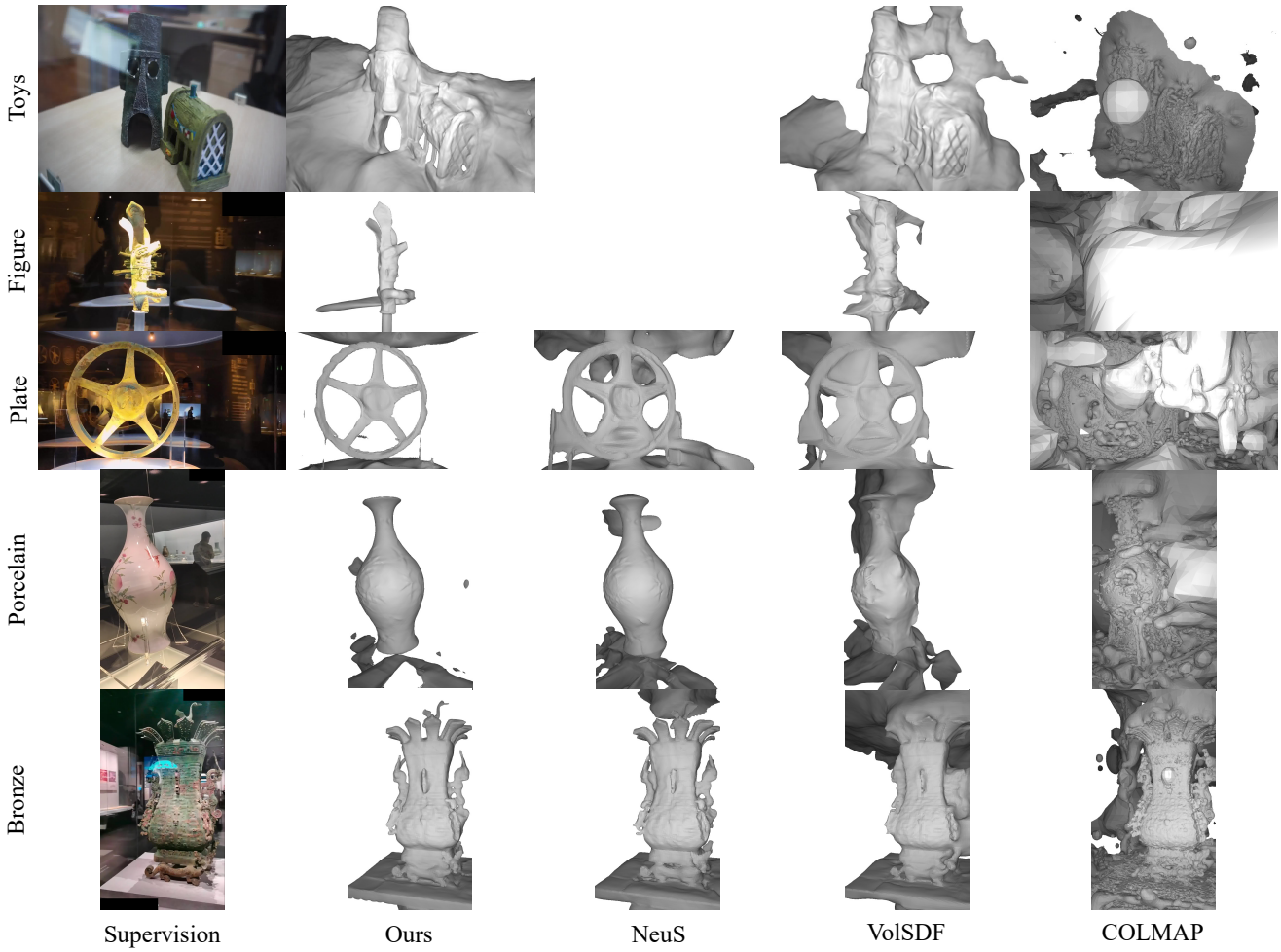
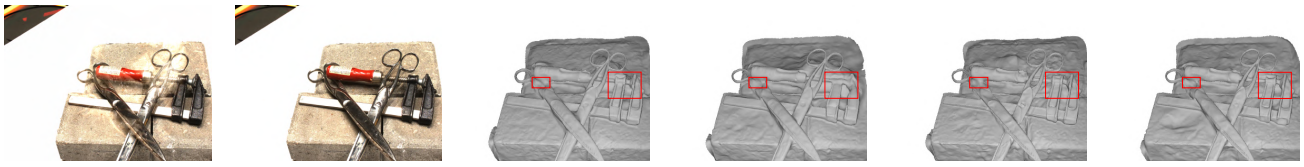


Figure 10. Qualitative comparisons on the real-world dataset. NeuS [45] fails to recover the meaningful object surfaces on 'Toys' and 'Figure'. Our method recovers target objects against HSR with physically correct surfaces, while other methods generate noisy results.



Chamfer Distance	2.29	3.69	1.96	1.89

Figure 11. Ablation study on a synthetic scene. When decomposing the scene without attributes (d_r and $\mathbf{n}_r$) of auxiliary planes, the performance degrades remarkably. The recovered target object surface loses fine details while disabling the volume density σ_r.

HSR scenes. The real-world HSR encodes more diverse ambiguous information than the synthetic HSR for recovering the target object surface. The results in Fig. 10 illustrate that our method generates better results containing thin structures of target objects when compared against state-of-the-art neural implicit methods.

4.5. Ablation Study

We conduct several ablation experiments to study the impact of different settings on the auxiliary plane module, including the volume density σ_r and the plane attributes (including the position d_r and the plane normal $\mathbf{n}_r$). Fig. 11 presents the results of each setting.

Effect of the plane attributes. For each camera ray in a view, we use MLPs to generate the volume density σ_r, and the attributes (d_r and $\mathbf{n}_r$) of an auxiliary plane. When we remove the attributes of the auxiliary plane and only utilize σ_r to generate the weight, the performance degrades drastically compared to the full model. Without the plane

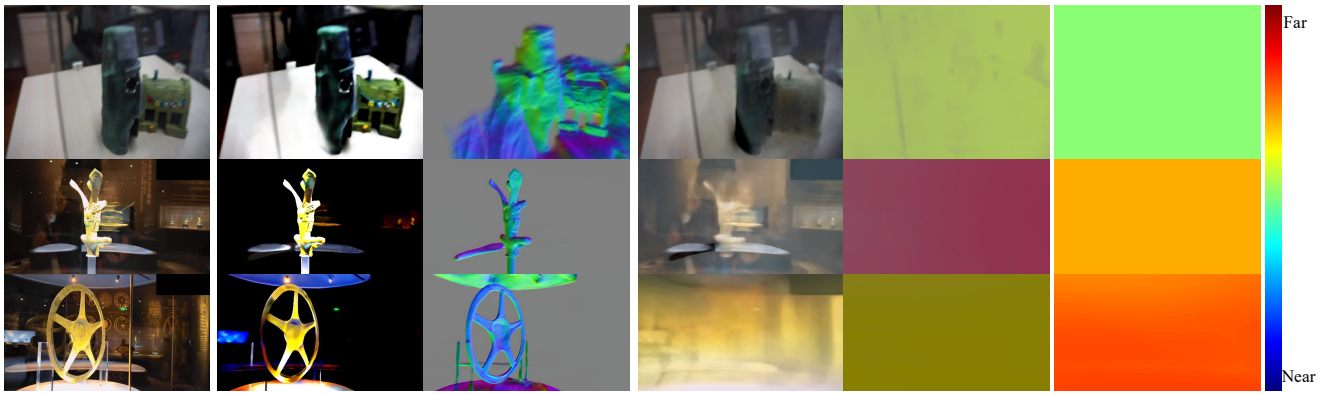


Figure 12. Components of NeuS-HSR. The rendered image of NeuS-HSR is decomposed into two appearances: the target object and the auxiliary plane. Target object appearances encode complete target objects with high sharpness, and auxiliary plane appearances enhance the content of HSR. The color bar represents the color map of plane position ('Near' means the auxiliary plane is close to the camera).

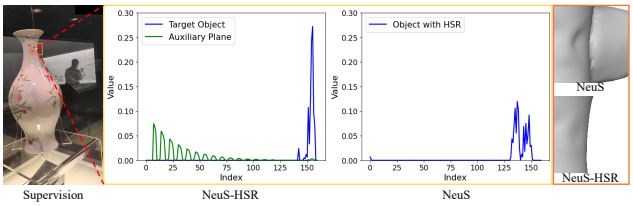


Figure 13. Rendering weights in a real-world HSR scene. NeuS-HSR enables NeuS to pay more attention to the target object against HSR, as the cropped meshes on the right show.

attributes, MLPs fail to implicitly trace the incident light to separate the target object appearance and the other part physically.

Effect of the volume density. To determine whether the volume density σ_r is necessary for recovering the object surface, we disable MLPs to output σ_r and adopt the same weights w as the object path to render two appearances. This operation introduces the ambiguity from two paths to the MLPs of predicting SDF, then produces a worse result than the full model. However, our model with this setting still achieves better performance than the baseline NeuS because of the robust auxiliary planes.

5. Discussion

Components. Our model consists of two parts: the target object and the auxiliary plane. Fig. 12 shows the components of each part. The target object appearances are faithfully enhanced and the HSR is captured by the auxiliary plane module. The surface normal and position of the auxiliary plane are adaptively learned by MLPs. The plane normals and positions on all camera rays of a view tend to be the same, which physically models a planar reflector.

Attention Analysis. In HSR scenes, to recover accurate target objects, our model should pay more attention to the object path rather than the plane path. As Fig. 13 shows, the rendering weights of the target object have a higher peak value and a more concentrated distribution than the weights of the auxiliary plane and NeuS. This demonstrates that the auxiliary plane module makes MLPs focus on the target object and then reduces the interference of HSR to achieve more accurate results.

Limitation. The proposed method inherits the ill-posed limitation from neural implicit methods of multi-view reconstruction. Due to the lack of priors, our model generates inaccurate geometry of target objects in unseen areas. A possible solution is introducing the symmetry of objects.

6. Conclusion

In this work, we have proposed a task of multi-view object reconstruction under the interference of HSR. To tackle this task, we present NeuS-HSR, a novel framework that recovers accurate 3D object surfaces against HSR. We propose decomposing scenes captured through glasses into the target object part and the auxiliary plane part for enhancing the target object by the auxiliary plane. We design an auxiliary plane module to physically generate the auxiliary plane appearance by using MLPs and the reflection transformation. Comprehensive experiments on both synthetic and real-world scenes illustrate that NeuS-HSR outperforms previous methods in quantitative reconstruction quality and visual inspection. Besides, the discussion explores the effectiveness of our decomposition in our task.

Acknowledgments

This work is supported by the National Key Research and Development Program of China Grant (No.2018AAA0100400), NSFC (No.61922046) and NSFC (No.62132012).

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 5

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 6

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 1

[4] Ya-Chu Chang, Chia-Ni Lu, Chia-Chi Cheng, and Wei-Chen Chiu. Single image reflection removal with edge guidance, reflection classifier, and recurrent decomposition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2033–2042, January 2021. 1

[5] Ziang Cheng, Hongdong Li, Yuta Asano, Yinqiang Zheng, and Imari Sato. Multi-view 3d reconstruction of a texture-less smooth surface of unknown generic reflectance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16226–16235, 2021. 2

[6] Ziang Cheng, Hongdong Li, Richard Hartley, Yinqiang Zheng, and Imari Sato. Diffeomorphic neural surface parameterization for 3d and reflectance acquisition. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[7] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 3

[8] Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *Proceedings of International Conference on Computer Vision (ICCV)*, volume 2, 1999. 1

[9] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W.H. Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5017–5026, October 2021. 1

[10] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[11] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[12] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25(361-369):2, 2016. 1

[13] Clement Godard, Peter Hedman, Wenbin Li, and Gabriel J Brostow. Multi-view reconstruction of highly specular surfaces in uncontrolled environments. In *2015 International Conference on 3D Vision*, pages 19–27. IEEE, 2015. 2

[14] Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable stereopsis: Meshes from multiple views using differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8635–8644, 2022. 2

[15] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409. IEEE, 2006. 2

[16] Ronald Goldman. *Matrices and transformations*. Graphics Gems, 1990. 2, 3, 4

[17] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning*, pages 3789–3799. PMLR, 2020. 5

[18] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022. 1, 2, 4

[19] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008. 2

[20] David S Immel, Michael F Cohen, and Donald P Greenberg. A radiosity method for non-diffuse environments. *Acm Siggraph Computer Graphics*, 20(4):133–142, 1986. 3

[21] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 3

[22] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2, 6

[23] Chenyang Lei and Qifeng Chen. Robust reflection removal with reflection-free flash-only cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14811–14820, 2021. 1

[24] Chenyang Lei, Xuhua Huang, Chenyang Qi, Yankun Zhao, Wenxiu Sun, Qiong Yan, and Qifeng Chen. A categorized reflection removal dataset with diverse real-world scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3040–3048, 2022. 1

[25] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 2

[26] Ming Liu, Jianan Pan, Zifei Yan, Wangmeng Zuo, and Lei Zhang. Adaptive network combination for single-image reflection removal: A domain generalization perspective. *arXiv preprint arXiv:2204.01505*, 2022. 1

[27] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[28] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 5

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3, 6

[30] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 3

[31] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3, 6

[32] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 3

[33] Sverker Rasmuson, Erik Sintorn, and Ulf Assarsson. Addressing the shape-radiance ambiguity in view-dependent radiance fields. *arXiv preprint arXiv:2203.01553*, 2022. 3

[34] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[35] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 1

[36] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 5, 6

[37] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006. 2

[38] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999. 2

[39] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[40] Binbin Song, Jiantao Zhou, and Haiwei Wu. Multi-stage curvature-guided network for progressive single image reflection removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1

[41] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[42] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 2

[43] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 3

[44] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 139–155. Springer, 2022. 3

[45] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 1, 2, 3, 5, 6, 7

[46] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. Hf-neus: Improved surface reconstruction using high-frequency details. In *Advances in Neural Information Processing Systems*, 2022. 3

[47] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 2

[48] Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6187–6197, 2022. 3

[49] Changchang Wu. Visualsfm: A visual structure from motion system. *http://www. cs. washington. edu/homes/ccwu/vsfm*, 2011. 2

[50] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. S3-nerf: Neural reflectance field from shading and shadow under a single viewpoint. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[51] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 3, 6

[52] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 1, 3, 5

[53] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[54] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 6

[55] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 5